RNA-seq: methodology and analysis

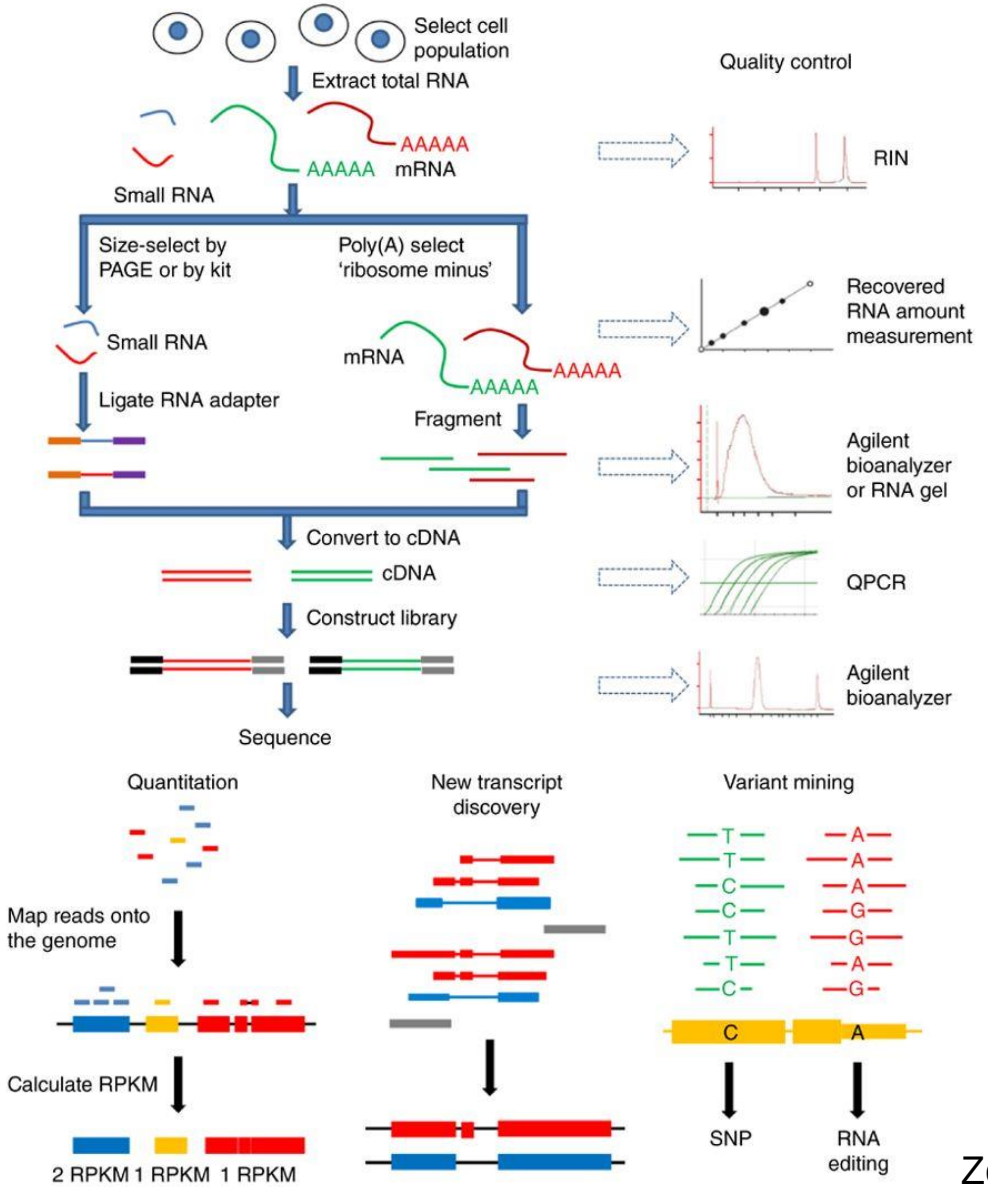
Jerry Zak

Expression profiling: microarray to sequencing

- **Microarray**
  - probe-based estimation of transcript abundance
  - blind to sequence, transcript structure (except exon arrays)
  - some probes work well, some less so (e.g. compare 2 probes for *Trp53bp2* in Allen Expression Atlas (human.brain-map.org/microarray/search/show?exact_match=false&search_term=TP53BP2&search_type=gene&page_num=0))

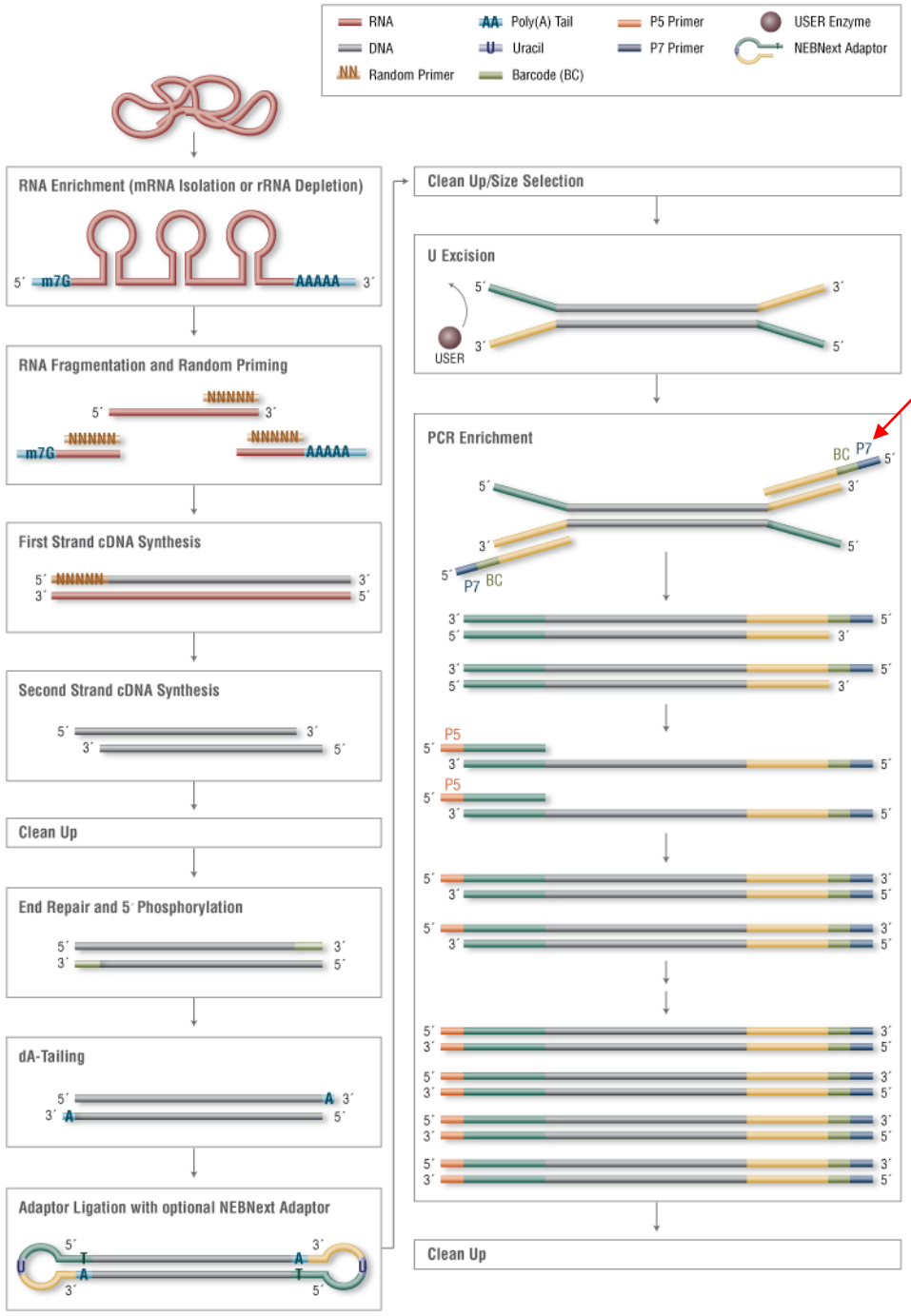
- **RNA sequencing**
  - expression profiling by high-throughput sequencing
  - can detect unannotated transcription, RNA species: discovery tool

RNAseq overview

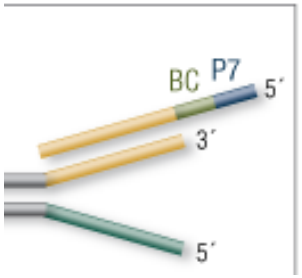


Zeng and Mortazavi 2012

Library prep for Illumina (example protocol)



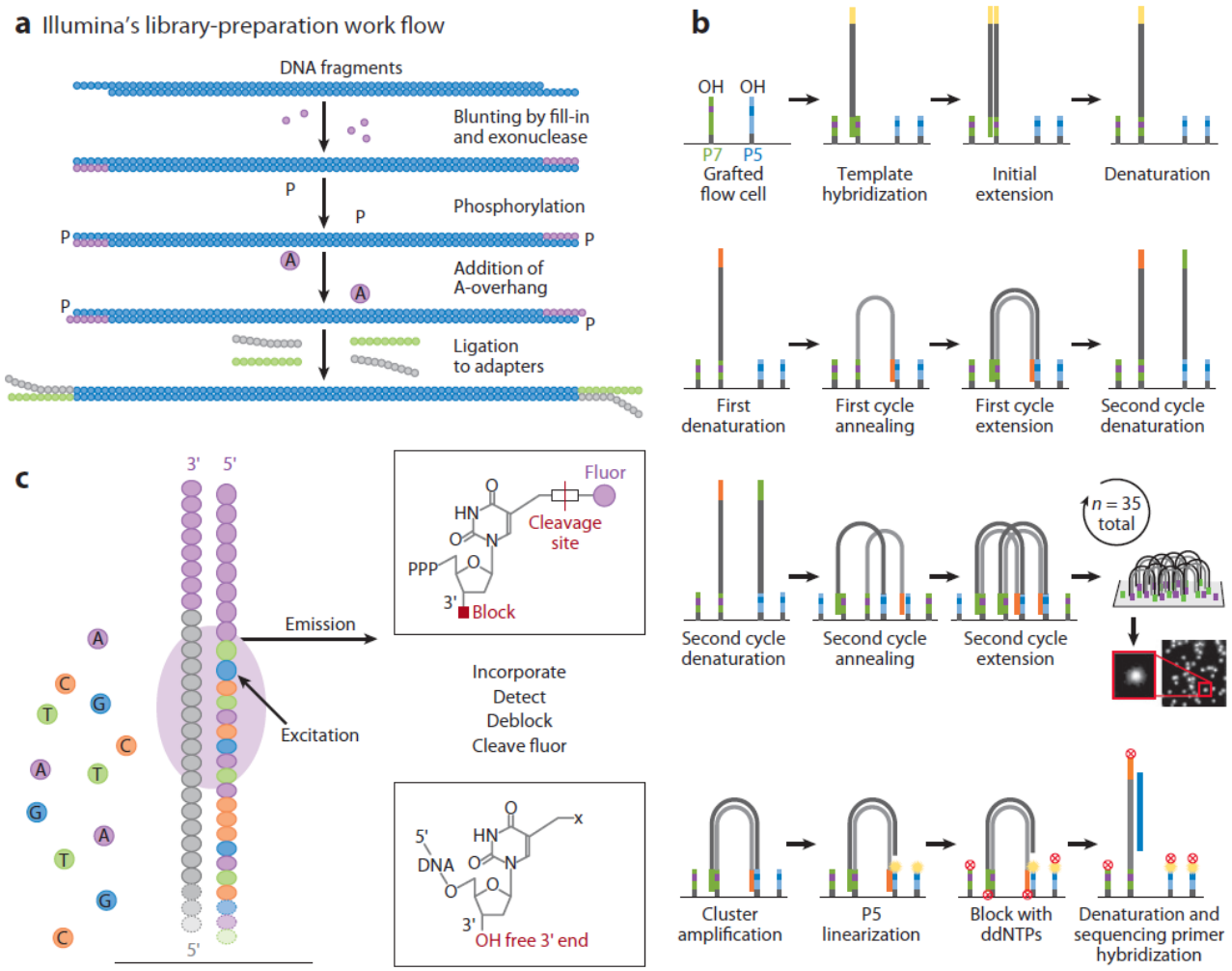
Adaptor – barcode – P7 primer



Neb.com

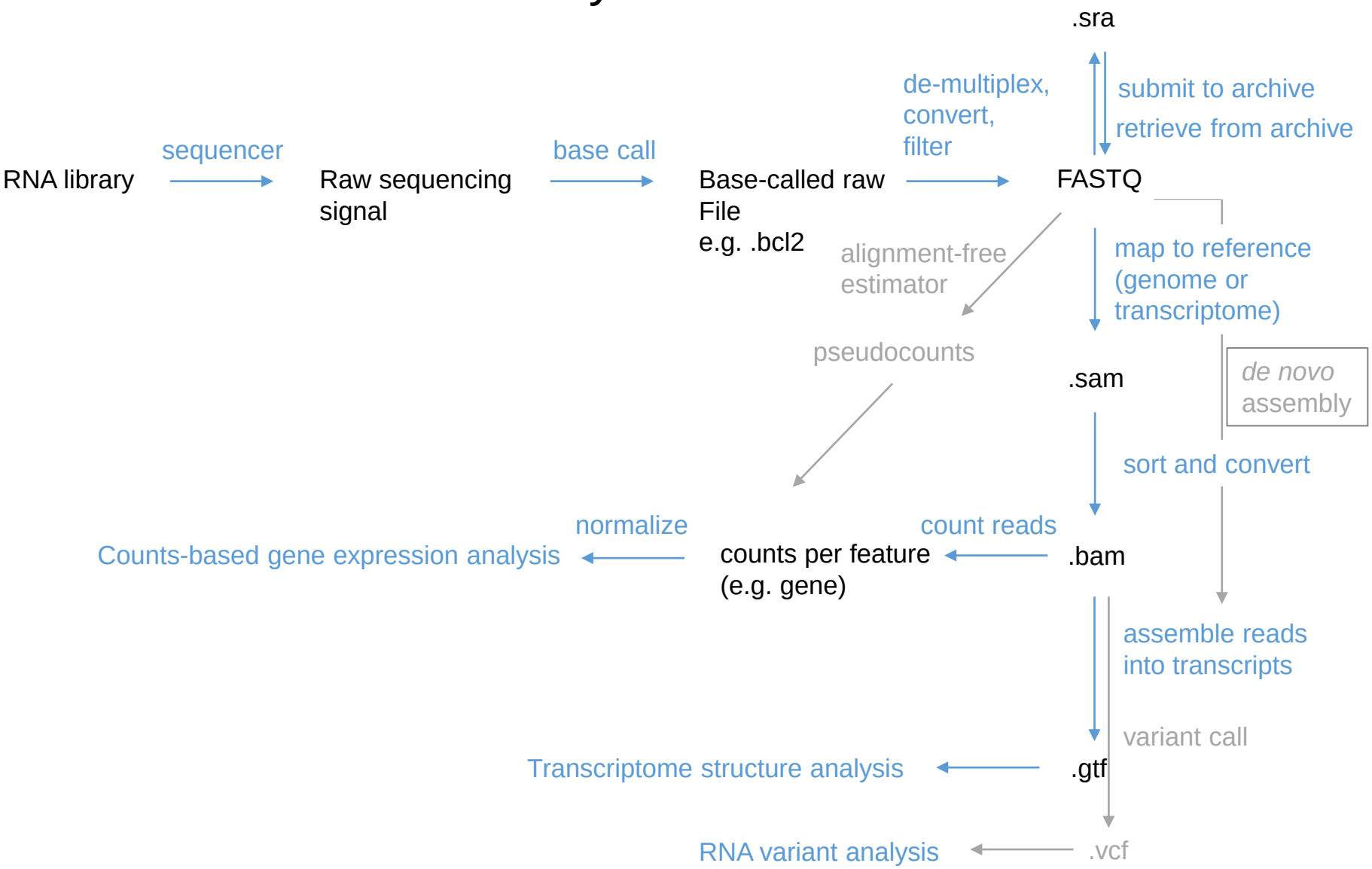
Illumina Sequencing by Synthesis Technology

- Illumina has a ca 75% market share in NGS (source: SA-Business Research & Consulting Group)
- Important alternative technologies: PacBio, Oxford Nanopore

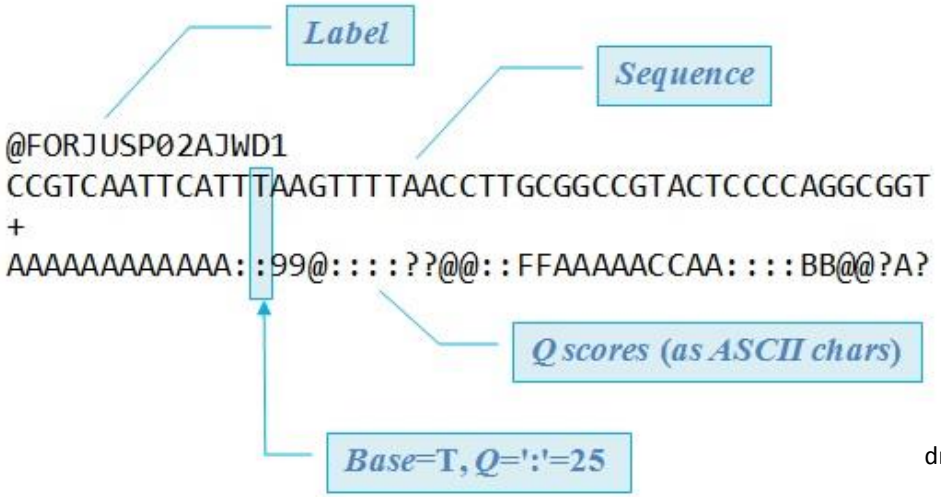


Mardis *Annu Rev An Chem* 2013

Analysis overview



Reminder: FASTQ format



drive5.com/usearch/manual/fastq_files.html

Illumina FASTQ

```
@NS500704:482:H3FCYBGX9:1:11101:20005:1037 1:N:0:CGCTCATT+AGGATAGG
CACCTNGGGCCCCCGGGCGGGGCCCTTCACCTNCATTGCGCCACGGCGGCTTTCGGACGAGCCCCTGACTCGCGC
+
AAAAA#EAE</EEAEEEEEEE//E/////AA/#<EEE/<<6/AE//EEE///EEEE6A///<//EEE/EEEAE//
```

For read tag info see https://help.basespace.illumina.com/articles/descriptive/fastq-files/

Illumina adapter sequences: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-08.pdf

RNA-seq Data sources

Processed

Gene Expression Omnibus

BioGPS, recount2, GTEX, cBioPortal, EMBL Expression Atlas…

Raw

Public

Sequence Read Archive

ArrayExpress

Restricted access

dbGaP GENOTYPES and PHENOTYPES

european genome-phenome archive

NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS

1. **Minimum Qualifications to Submit a dbGaP Project Request as a PI**. Investigators must be permanent employees of their institution at a level equivalent to a tenure-track professor or senior scientist with responsibilities that most likely include laboratory administration and oversight. Laboratory staff and trainees such as graduate students, and postdoctoral fellows are not permitted to submit project requests.

2. **Supplemental Documentation**. Some datasets require local Institutional Review Board (IRB) approval for use, as noted on the dbGaP study page. Other types of documentation may be required as described on the study page and/or Data Use Certification for the study. Evidence of IRB approval and other documentation can be uploaded as a PDF during the application process.

3. **Accessing Additional Datasets After Initial Approval.** Investigators who would like to access additional dataset(s) for use in an existing approved project should (1) revise the existing

Demonstration: interacting with data

Programmatic access

Web access

Exercises I:
1) What read length was used in the sequencing study *A Distinct Population of Thirst-Associated Preoptic Neurons Encodes an Aversive Motivational Drive*?

2) Using bash, print the first 5 ‘spots’ of the first RNA-seq run from the study above

HW0 Now print the first 100 ‘spots’ of the first 5 runs (alphabetically) of the study each into a separate file and quantify how many times the nucleotide adenosine occurs in each one. How does this compare to the other nucleotides? What are the possible reasons?

Focus on *Applied* Bioinformatics

How to approach practical research questions regarding RNA expression?

- Do not jump into raw data analysis by default, utilise processed data resources where applicable
- Sophistication of methodology should match scale of the problem

Case 1: A knockout mouse line deficient in gene G has unexpected male infertility. Could this gene play a role in sperm development or function?

Approach: Check publicly available expression data like GTEX, BioGPS, The Human Protein Atlas for testis expression i.e. **processed data.**

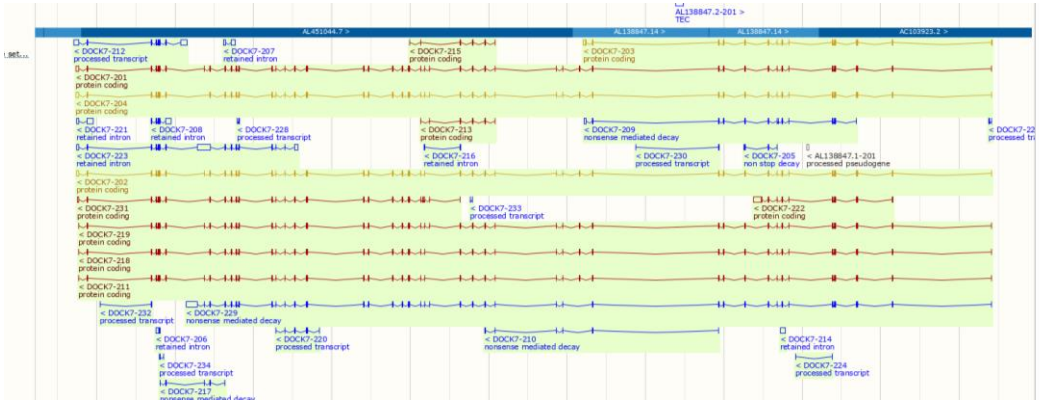
Case 2: You identified a novel protein that induces cell cycle arrest in cell culture. Is there any evidence that its gene is a tumor suppressor in humans?

Approach: Mine publicly available cancer genomic data using a **processed data** portal such as cBioPortal.

Focus on *Applied* Bioinformatics (2)

Case 3: A Western blot shows multiple bands with a single antibody, and the bands disappear when the mRNA is knocked down using siRNA. Could alternative splicing be responsible?

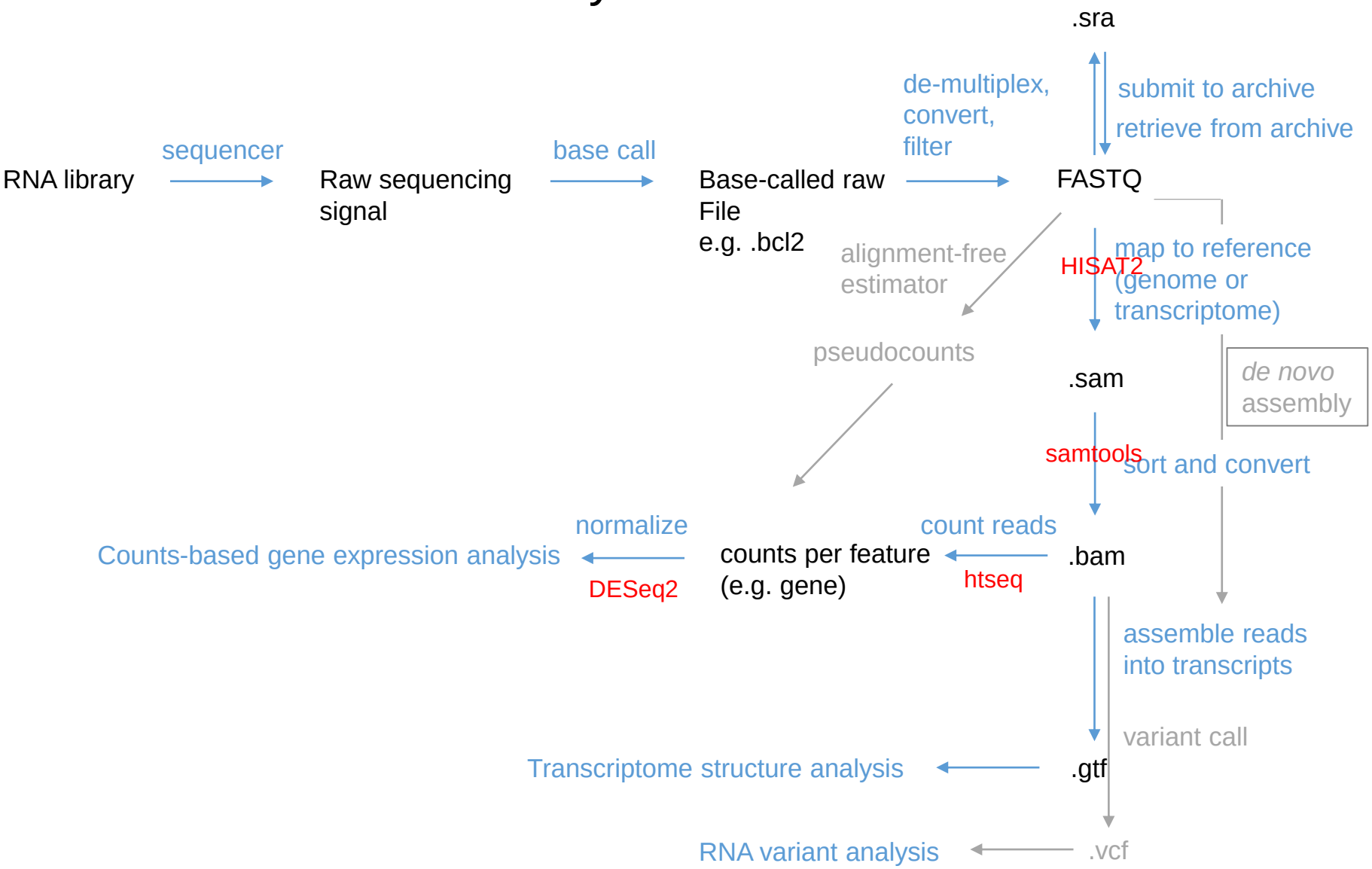
Approach: Investigate the annotated transcripts on Ensembl and UCSC Gene Browser. Explore transcript-level expression in GTEX.



Case 4: You hypothesize that endogenous retroviruse (ERV) expression correlates with immune cell infiltration in human cancer.

Approach: Re-map raw RNA-seq to human genome with improved HERV annotations. Requires extensive statistical design.

Analysis overview



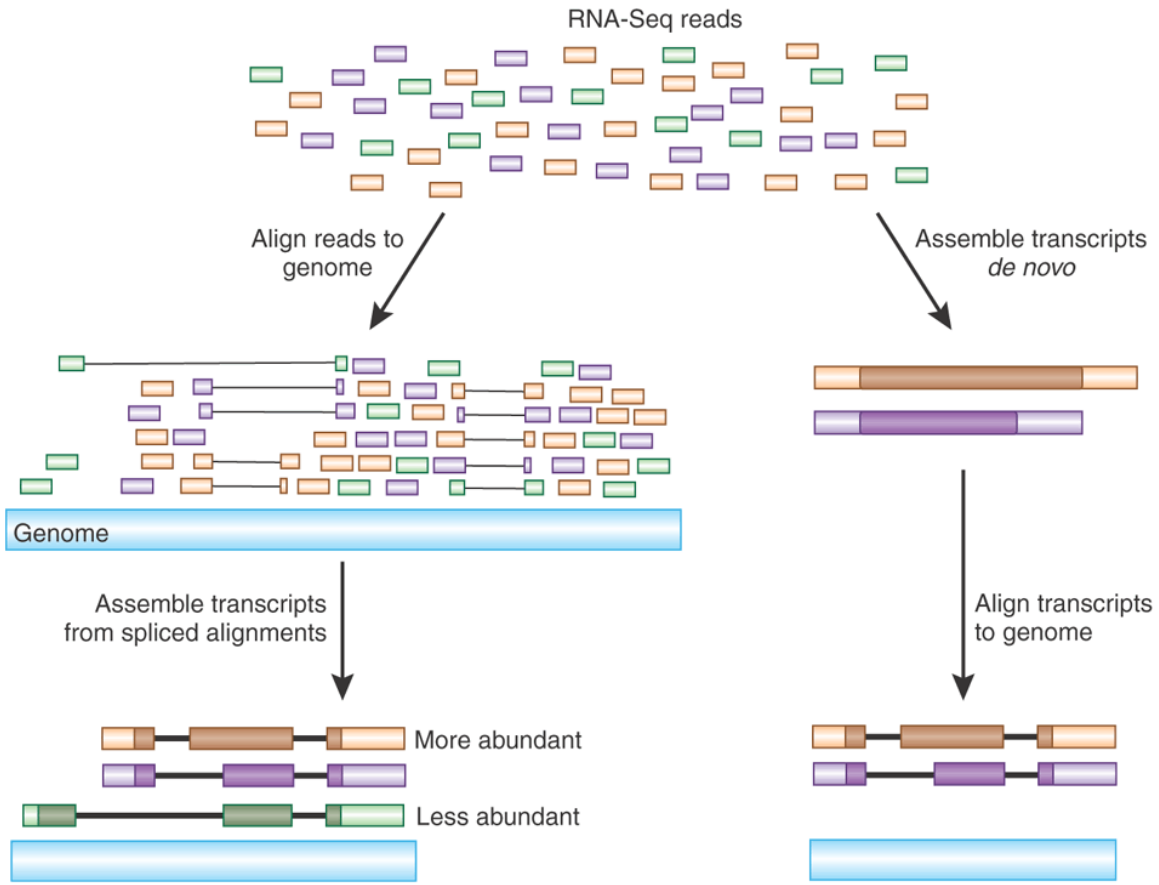
RNA-seq analysis is still evolving

- Check out the literature for new methods and benchmarking studies

e.g. correlation with qPCR-validated dataset (ERCC) : differential expression performance

Sahraienan 2017

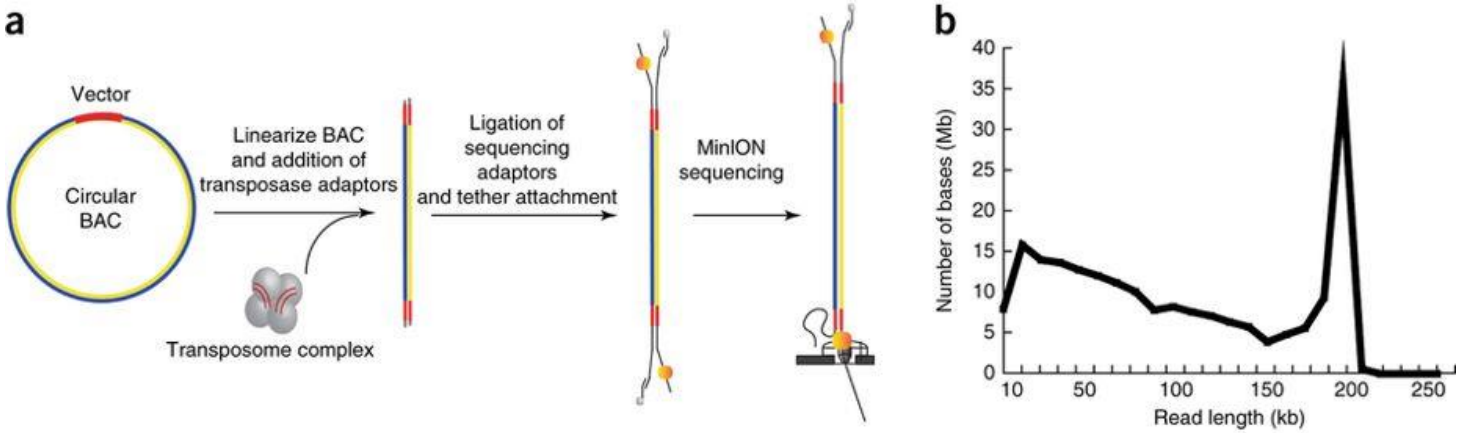
Mapping to a reference



http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

Single-end, paired-end & full-length reads

Illumina.com



Jain et al, *Nat Biotech* 2018

Demonstration: mapping

Exercises II:
1) It is common for RNA-seq reads to be contaminated with leftover adapter sequences. Test if the file SRR5454079_1.fastq contains any i7 adapter sequences ATCACGAC or ACAGTGGT.
1-2) What would be a proper way to determine if there is significant barcode contamination?

HW1: Map SRR5454079 to the human genome using HISAT2 (or chromosome 20 if you have <8GB RAM).
HW2: Check the first 5 reads contained in the FASTQ files. Which chromosome did they align to?
HW3: Consult the samtools documentation then convert the output sam file into a sorted bam file.